DOCUMENT RESUME

ED 292 059 CS 009 062

AUTHOR Thompson, Charles L.; And Others
TITLE Speech Recognition Technology: An Application to Beginning Reading Instruction. Technical Report.
INSTITUTION Educational Technology Center, Cambridge, MA.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Oct 85
CONTRACT NIE-400-83-0041
NOTE 45p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Beginning Reading; *Computer Assisted Instruction; Elementary Education; *Reading Instruction; *Reading Programs; Science Programs; Speech Synthesizers; *Technological Advancement
IDENTIFIERS *Computer Speech Recognition; Software Testing

ABSTRACT

Noting that the recent development of reliable, high-performance, low-cost speech recognizers--devices that can distinguish among spoken words--holds potential for education, such as early reading instruction, this technical report describes a study which investigated two principal questions: (1) Does an inexpensive, microcomputer-based speech recognizer perform reliably enough on young children's speech to permit application to reading instruction?; and (2) What are the main human factors attending such use? The Dragon System Mark II Isolated Word Speech Recognizer was used in the study, which included four stages. The first phase took place in June 1984 and involved 17 kindergartners; the second phase took place in November 1984 and involved 7 kindergartners and 8 first graders; the third phase took place in late December and involved 10 kindergartners; and the fourth phase took place in August 1985 and involved 6 students who had completed kindergarten and were about to enter first grade. The results of the study indicated that speech recognition technology holds potential for such educational applications as beginning reading instruction. Findings also suggest that human factors, such as microphone handling, responses to recognition errors, responses to prompts and remarks, and need for adult supervision are crucial ingredients in the effective application of speech recognition technology in education. (Seven tables of data are included and a short bibliography is attached.) (NH)

SPEECH RECOGNITION TECHNOLOGY:

AN APPLICATION TO BEGINNING READING INSTRUCTION

Technical Report

October 1985

Educational Technology Center
Harvard Graduate School of Education
337 Gutman Library Appian Way Cambridge MA02138

SPEECH RECOGNITION TECHNOLOGY:

AN APPLICATION TO BEGINNING READING INSTRUCTION

Education Development Center, Inc.
Newton, Massachusetts

October 1985

Charles L. Thompson
Beth Wilson
Philip P. Zodhiates
Ilene Kantrov

Preparation of this research report was supported in part by the National Institute of Education (contract # NIE 400-83-0041). Opinions expressed herein are not necessarily shared by NIE and do not represent Institute policy.

ACKNOWLEDGMENTS

Dragon Systems, Inc., not only provided the speech recognizer used in this study, but also served as an important technical resource on a wide variety of issues. Janet M. Baker, President of Dragon, reviewed an earlier draft of this report, correcting technical inaccuracies and suggesting ways of analyzing data. Mark Sidell, principal hardware engineer at Dragon, designed the prototype printed circuit board that we used, and made hardware adjustments based on our field testing. Jed Roberts and David Pinto, research scientists at Dragon, attended portions of the field testing, carried out software adjustments to the recognizer, and helped us think about the meaning of some of our data.

John H. McCloskey, senior programmer at Education Development Center, Inc., (EDC) was responsible for programming the reading instruction software and revising it based on the field testing. David Nelson, EDC s Director of Media Services, conducted the videotaping of the test sessions.

We are also grateful to the staff and students of the Phillips School in Watertown, Massachusetts, the Lincoln-Eliot School in Newton, Massachusetts, and the Plowshares Childcare Program in Newton, Massachusetts, who participated in the field tests. In addition, the Watertown school system kindly loaned us videotaping equipment, which allowed us to record the testing, and the Education Collaborative for Greater Boston, Inc., provided space where the testing was conducted.

TABLE OF CONTENTS

Speech synthesizers--devices that produce spoken language from digital code--have become familiar adjuncts to microcomputers and have been built into some computers intended for educational use. Speech recognizers--devices that can distinguish among spoken words--are a more recent phenomenon. Until recently such devices had been either too expensive or too poor in performance to support widespread educational use. The highest quality recognizers are usually priced in the tens of thousands of dollars, while the more affordable devices often have accuracy levels well below the 97 percent experts suggest is necessary for satisfactory use.

Recently, however, the price-performance picture has changed dramatically. Dragon Systems, a Newton, Massachusetts-based speech technology research and development firm, has introduced speech recognition software supported by a printed circuit board that can be priced well below $500 and that performs with 99.3% accuracy, according to a standard test of isolated word recognition. Though Dragon's device is limited to small active vocabularies of isolated words (sixteen to thirty-two words at a time), its total vocabulary is constrained only by available memory. Its development signals the advent of high-performance, low-cost speech recognition and places this technological potential within the reach of developers of educational software.

A reliable speech recognition capability available at a low price appears to hold considerable potential for education, especially for

educational tasks in which speech is essential, such as early reading instruction. In this study we have used the Dragon Systems Mark II Isolated Word Speech Recognizer to investigate two principal questions:

--Does an inexpensive, microcomputer-based speech recognizer perform reliably enough on young children's speech to permit application to reading instruction?

--What are the main human factors attending such use?

It should be emphasized that our research has focused almost exclusively on technical and human factors issues, rather than on pedagogical or psychological questions.

The Mark II is mainly in software form. To operate in a standard Apple II or IIe microcomputer, the Mark II's only requirements are the printed circuit board designed by Dragon and an inexpensive commercially available microphone. No other additions to the standard one-drive configuration are required.

The system is speaker-dependent, requiring each user to train the recognizer by giving a few samples of his or her pronunciation of each word to be recognized. The Mark II analyzes these samples and constructs templates against which to compare subsequent utterances. In some applications speaker dependence is considered a disadvantage because it requires each new user to train the system and uses only that speaker's own templates for future recognition. Our preliminary research, however, indicated that speaker dependence might in fact be an asset with the highly variable pronunciation of young children's speech.

For use with beginning readers, of course, text display does not suffice as the sole or even primary mode of output. The obvious alternatives are graphics, music, and speech synthesis. In an area such as

reading instruction, speech output is especially important. Moreover, the speech output must be of sufficient quality to introduce new words to students, not merely to produce recognizable utterances of words already known. In this study we have experimented with two different methods of producing high quality speech output from digital code. Both rely on compression of actual recorded speech rather than on text-to-speech synthesis.

Prototype versions of software under development by EDC were used in the study to provide the reading tasks and the graphics and computer-generated music used to prompt and reward reading performance.

BACKGROUND ON SPEECH TECHNOLOGY

Existing speech recognition systems differ along a number of dimensions: speaker independence vs. speaker dependence, isolated word vs. continuous speech recognition capability, and vocabulary capacity.

Most currently available systems are speaker-dependent, that is, they must be trained to respond to individual speakers. Because of the acoustic variability in phoneme production among different speakers, these systems are more reliable than speaker-independent ones, which are intended to work for all users. Speaker-dependent devices work by having each user provide a few samples of his or her pronunciation of the specific words to be recognized. The recognizer samples the voice waveforms thousands of times per second, digitizing them and computing information on the frequency and temporal characteristics of the speech. From this information the device constructs templates of the sample utterances and stores them in memory. Each speaker must repeat each word several times so that the templates can adequately capture the variability of that

particular voice. Some systems, including the Mark II, average together all sample utterances (or tokens) to construct one composite template; many others store each token as a distinct word. Although insufficient training data is a major cause of speech recognition errors, training can be tiresome for users, and therefore most systems currently suggest only a few tokens of each word. In addition, most systems are not computationally capable of handling more training data. High-performance speaker-independent systems typically require a very large data base, constructing their templates from hundreds or thousands of different speakers.

Once usable templates have been constructed, the recognizer compares new utterances to them, using statistical algorithms that describe the acoustic parameters of words. Some systems are word-based: they compare all of the information extracted from a new utterance to the stored templates and find the best statistical match between the utterance and one of the templates. Other systems attempt to segment the speech waveform into acoustically distinct regions and to compare selected portions, rather than the entire waveform. Although this feature-based method reduces the amount of computation required in template matching, and thus is cheaper and more efficient, it often sacrifices a degree of accuracy. Some recognizers use a process called dynamic time warping in which the waveform of an incoming word is figuratively stretched or compressed to match a stored template. This technique gives the recognizer added flexibility in compensating for variations in user pronunciation.

Speech recognition systems also differ on the basis of whether they recognize single isolated words or continuous speech. Some of the most sophisticated--and most expensive--systems are capable of recognizing

continuous speech with total vocabularies of up to 500 words. Users develop an application-specific grammar that predetermines groups of meaningful word sounds. Spoken words are recognized as valid only if they conform to this predefined grammar. This reduction in the number of possible word combinations increases accuracy and conserves computation. Not surprisingly, the cost of a system with this degree of sophistication places it well out of reach for any widespread educational application. And even these recognizers designed to work with continuous speech perform better with isolated words, so much greater is the acoustic variation of words spoken in continuous speech than of words spoken individually. Most speech recognizers have a total vocabulary--say, 100 words--and a smaller active vocabulary--say, 35 words--which they are actively "listening for" at a given time. These words are stored in random-access memory where they can be quickly and easily available for matching with incoming utterances. The majority of systems recognize only isolated words, and the template matching process begins as soon as the end of a word is detected. Continuous speech systems may begin template matching before the word is completed. In either case, as vocabulary size increases, so do storage and computing requirements, and thus cost.

A major factor in the usefulness of any system is its accuracy in recognizing the vocabulary on which it has been trained. The primary means of assessing reliability is to look at how well a system performs with respect to two principal kinds of errors: substitution errors, or the mistaking of one word for another; and rejection errors, the refusal to recognize a valid utterance. These two types of errors have a partly reciprocal relationship; if the system is designed to make fewer rejection errors (rejecting fewer "correctly" pronounced words), it is therefore

likely to make more substitution errors, that is, recognizin words incorrectly. A third kind of error, insertion--the recognition of background noise or other speech as a valid word--may also be considered in determining accuracy levels.

One of the factors that can compromise accuracy has already been mentioned: variability, both among different speakers and among the same speaker's utterances. Aside from the acoustic characteristics of phoneme production, this variability stems from changes in voice volume and from nonspeech-related sounds such as tongue clicks, breath sounds, and inadvertent "um's" and "er's." Some individuals seem to have more variable speech than others, and conventional wisdom within the field of speech technology holds that approximately 80 percent of recognition errors occur among 20 percent of users. An additional difficulty is posed by the noise environment in which the recognizer is being used. This may include noise from machinery, movement, and electronic noise, as well as background voices. In fact, competing background voices are usually more disruptive than other environmental noise. Some recognizers use a calibration system so that they can be adjusted for different noise conditions; others are set at a fixed level.

The quality of the microphone also influences reliability. A system intended to function in a noisy environment has to screen out a great deal of background noise. In the process it may also screen out some valid utterances and thus produce a higher rate of rejections. On the other hand, screening out less background noise will likely produce a higher rate of substitutions (for example, recognizing "car" for "tar") and insertions (for example, recognizing background noise as a word). In some environments a high performance directional headset microphone is neces-

sary to obtain an adequate level of accuracy.

Until now, applications of speech recognition technology have been limited mainly to business and industrial settings. For example, speech recognizers have been used for office automation, automatic telephone transactions, and inventory control and quality assurance inspections in factories. In this study we explore the potential of speech recognition technology in education.

The Dragon Mark II, by combining some strategies of word-based template matching with more feature-based recognition, and by using more sophisticated algorithms based on stochastic processing models of language, has been able to achieve a 99.3 percent accuracy level\1/ with greatly reduced computational requirements. Its vocabulary, though small (16-32 words active, 200 total), is more than adequate for use in many educational tasks, such as beginning reading instruction. Speech recognizers like the Mark II may enable microcomputers to supplement the work of the teacher or other skilled reader in confirming and correcting the efforts of young learners.

Synthesized speech output, though not the focus of investigation in this study, deserves brief mention here because it provides a necessary and important complement to the speech recognizer. Speech output, especially for students who cannot yet read, becomes the primary means of providing instruction and feedback to the child. In our study we tested two different methods of producing speech output, both of which relied on compression of actual recorded speech. Initially, we employed a speech synthesis technique that sampled recorded speech at a high rate. Though this method produced natural-sounding speech that was easily understood by young users, it had the disadvantage of requiring large quantities of disk

storage space and multiple drives. Consequently, we replaced the extra disk drives with a Texas Instruments speech synthesis chip; this more economical method used linear predictive coding coefficients which were computed from recorded live speech. We wanted to determine whether its somewhat less natural sound would still be adequately understood by kindergarten and first grade students.

METHODS

Field testing of our prototype speech recognition system was carried out in four phases.

Phase 1:

The first field testing occurred on two consecutive days in June 1984, in a Watertown, Massachusetts, public elementary school. The participants, 17 kindergarteners--12 boys and 5 girls--were nominated by their teachers and were given the option of participating in the study. Standard school district policies on parental and student consent were followed. Teachers were asked to choose students from a range of ability levels; according to teacher classification, the test group comprised 5 low, 5 average, and 7 high ability readers. Seven students were 5 years old, nine were 6 years old, and one was 7 years old.

Students were taken from their regular classrooms and conducted to another room in the school where the system had been set up. To reduce any anxiety associated with the experiment, students were taken in pairs. The entire testing procedure for each student required from fifteen to twenty minutes.

Upon entering the room students were seated together in front of an Apple IIe microcomputer equipped with an inexpensive, commercially avail-

able microphone. On the first day children held the microphone in their hands; on the second day the microphone was placed in a stand to determine whether a stationary position would increase the accuracy of the speech recognizer. The testing was conducted by three EDC staff members: one, referred to here as the experimenter, assisted the students as they took turns completing the series of activities involved in each test session; another collected observational data; and a third videotaped the sessions.

First, students were pretested on the eight words used in the prototype program, plus two words that controlled the menu, yes and no. The students were shown these ten words on index cards and asked to read them.

Following the pretest the experimenter explained to students that they would be learning some new words by reading a story and playing a game with the computer. They were told that the microphone would help the computer to hear them and were instructed in how to hold the microphone. Background noise levels were monitored and recorded with the use of a professional quality sound meter.

The system was then booted up, and children began to receive their instructions from the synthesized speech output, provided in this phase by the higher quality speech compression method. The training was begun as the speech output told participants that the computer needed to know how they said some special words. Beginning with the first word, the speech output instructed students to "please say _____." After each utterance the computer either signaled acceptance by moving the word on the screen, completing part of the graphic, and playing a musical tone, or instructed the child to say the word again. If, after four utterances were accepted, the system was able to construct a usable template, the child was rewarded with a musical phrase. In the case of a bad template, the child was

instructed, "Sorry, we need to do that one again," and the training was repeated for that word. Once all the templates were formed, the recognizer checked them by asking the child for one more sample of each word. In some instances a second round of retraining was necessary.

The testing continued with the reading of the story. The computer produced speech output for the text as it appeared on the screen, accompanied by graphic illustrations. Target words were shown in large letters, and the speech output instructed the child to "read the big word." Then, for each target word, the program would pause, giving the child the opportunity to read the word aloud. The recognizer had to match the child's utterance against the stored templates to determine whether the child had responded correctly. If the child supplied the correct word, he or she was rewarded with a graphic display and a musical phrase. If the child made a mistake, the computer repeated the sentence stem from the story and again paused for the child to read the word. If the child did not produce the correct word on the third trial, the computer provided the answer.

The final component of the prototype software was a game composed of a 3x3 matrix containing eight words and an empty center square. By reading the word from a box that bordered the empty square, children were able to move that word into the empty box. The objective was to move the word which began in the lower left-hand corner to the upper right-hand corner. The game required not only reading, but also a certain amount of strategy and planning, as only the words next to the empty square could be moved. At this early stage of the project the instructions for the game were not available by means of speech output and had to be provided by the experimenter. The experimenter also provided help when needed in reading

the words and planning strategy as children progressed through the game.

In the game, as in the story, the recognizer had to match the child's utterances against stored templates. Whereas in the story the recognizer listened for the one correct target word for each frame and rejected all other words, in the game its task was more difficult. It had to listen for and discriminate among all ten words at once. In both the story and the game segments, the possible recognition errors were rejection of correct responses supplied by the child or acceptance of incorrect responses. In addition, the recognizer could fail to respond at all because of problems related to microphone position or voice volume, or it could accept extraneous human or nonhuman environmental noise.

At the conclusion of the session, students were posttested to determine how many words they had learned. In turn, they were presented with the words on index cards as in the pretest. They were also asked a few questions about their experience with the system.

In addition, we collected structured observational data on three types of recognizer error: rejection (a valid utterance of a trained vocabulary word is not recognized as such), substitution (one word is recognized for another), and insertion (background or extraneous speaker noise is recognized as a valid utterance). For each error we noted when it occurred in the sequence and what might have caused it. We hypothesized that errors would be caused by students speaking too softly or too loudly, by their holding the microphone too far from the mouth, by extraneous background noise or speech, or by other factors we could not anticipate but hoped to identify through observation.

Through semi-structured observation, we also collected data on human factors associated with the educational use of speech technology:

prompting, microphone handling, and response to recognizer error. We noted whether children were able to comprehend and respond to the synthesized speech output and whether they were able to use the microphone appropriately and modulate their voices effectively enough to train the machine. We also observed their responses to machine errors, including "no-hears" and misrecognitions, and we noted the level of instructional support and prompting required for microphone use, for interpreting feedback from the recognizer, and for persisting with the task, as well as for help in word reading and generating game-related strategies.

In addition to these observations, we asked students about their experience with the speech recognition system. Our questions were designed to determine what they liked most and least, found easiest or hardest, or would like to see changed. In addition, we wanted to know whether they had any prior experience using computers and whether they would like to use the speech recognition system again in the future.

Finally, we videotaped all the testing sessions and interviews in Phase 1, both for purposes of simple documentation of the experiment and for possible future use in a tape designed to introduce speech recognition applications and issues to New England educators.

Phase 2:

In November 1984, we carried out a second round of testing, again in the same Watertown public elementary school. Seven kindergarteners and eight first graders tested the speech recognition system over a period of three consecutive days. Of the kindergarteners, three were girls and four were boys; two were 4 years old, and five were 5 years old. Of the first-graders, five were girls and three were boys; seven were 6 years old, and one was 7 years old. As in Phase 1 they were nominated by their teachers

who rated their reading abilities. Of the kindergarteners, one was rated as low in ability, two as average, three as high, and one was not rated. Of the first graders, two were rated as low, two as average, and four as high in ability.

Phase 2 included certain modifications in the speech recognition system itself and in the sequence of reading activities. Speech output was produced with the speech synthesis microchip from Texas Instruments. The major advantage of this modification was to reduce the required number of disk drives from four to one. Its one disadvantage was to yield speech output of slightly lesser quality than the previous method. To facilitate training, the speech recognition system was adjusted to require a pause of a fraction of a second before each sample utterance; this was intended to prevent acceptance of only fragments of words inadvertently spoken over speech output prompts. In addition, adjustments were made in both hardware and software to improve the recognizer's performance. These adjustments included finetuning of the software parameters for recognition as well as changes in the gain setting and the volume threshold for incoming utterances. Adjustment of the gain affects sound amplification and therefore the recognizer's responses to background versus foreground noise. A separate hardware adjustment raised the upper volume level which the recognizer would accept. The microphone was the same one used in Phase 1, and it was again placed in a stand in front of the computer.

In this phase a second set of eight words was added, making a total of eighteen possible words. Students were pretested on all eighteen words. Each student trained the system on the first set of words and then read the story or played a game. For nine students this process was repeated with the second set of words, but with the difference that for

both story and game the two sets of words were randomly intermingled. Because of time constraints, six students did only the first set. A full session including both sets of words required between thirty and forty minutes.

Training in Phase 2 included some different ways of prompting students to repeat the words. On the first three vocabulary words the machine used a 5-second delay after each utterance to allow the child to repeat the word again. If the child failed to do so, the computer prompted again by asking the child for another utterance. On subsequent words a 10-second delay was used to allow the child even more time to anticipate the prompt and thereby move through the training more quickly and eliminate the boredom of hearing the prompt again and again.

Second, if after two attempts at training a particular word, the recognizer still did not have a good template, the word was moved to the end of the list. This was done to eliminate boredom associated with having to repeat the same word many times in succession. Finally, for the observer's information the computer provided different output tones for words that were too loud or too soft, as well as for those that were in the appropriate volume range but rejected for other reasons.

Prompts were also modified in the story. If children initially supplied an incorrect word they were given a "watch this and then try again" prompt followed by a graphic clue. If they still failed to supply the correct word, the computer read them the answer.

A new "concentration" type game was added in this phase. Students had to match the shapes behind the words in a 3x3 matrix. The shapes were exposed by reading the words. Eventually, as matches were made, the picture of a monkey emerged. Some children played this game, some played

the game described in Phase 1, and some played both.

Pre- and posttesting were conducted, and other structured and semi-structured observational data were collected as in Phase 1. In addition, recognizer error data were recorded for all portions of the prototype software.

Phase 3:

A third round of testing was conducted in late December in a public elementary school of Newton, Massachusetts. Of the ten kindergarteners who participated, five were boys and five were girls. Three were rated by their teachers as low ability readers, two as average, and five as high. Two were 4 years old, six were 5 years old, and two were 6 years old.

The major innovation in this phase was the use of a headset microphone. Though this microphone was less expensive and of lesser quality than the one used in Phases 1 and 2, it was hypothesized that it would improve the performance of the recognizer by remaining a consistent distance from the child's mouth at all times. To facilitate training, synthesized speech was used to prompt students when their responses were too loud or too soft. The brief pause required before each training utterance was retained from Phase 2, but shortened. In addition, a number of further adjustments to the recognition software were aimed at improving its performance.

The third field test used the same sequence of activities as in Phase 2 and the same eighteen words. The major addition to the program was a game of tic-tac-toe in which the computer played against the child. Other modifications included the addition of several prompts and directions in each segment of the program. At the beginning of the training, the compu-

ter instructed children explicitly and demonstrated how they should time their utterances by watching the movement of the word across the screen. The computer also provided feedback of "louder please," "I can't hear you," and "That's too loud" for utterances that were outside the appropriate volume range for the recognizer. In the story, initial wrong responses were followed by a "try again" prompt, and second wrong responses were followed by the "watch this and then try again" prompt with a picture cue. If the child failed to respond at all, the computer waited 10 seconds and then delivered the latter prompt and cue. At the beginning of each game the computer began by giving instructions and a demonstration of how to play. During the games, if the child made an incorrect response or failed to respond, the computer repeated the directions or suggested a correct response.

Phase 4:

A fourth and final round of testing took place in early August 1985 with students enrolled in a private summer program in Newton, Massachusetts. Of the six students who participated, four were boys and two were girls; all had completed kindergarten and were about to enter first grade in September. Five of the children were six years old, and one was five years old. In this instance we did not receive teacher ratings of students' reading abilities.

The major innovation in Phase 4 was the use of a high quality, noise-cancelling, headset microphone. All other components of the speech recognition system, with the exception of an adjustment to e gain setting, remained the same as in Phase 3. In previous rounds of testing we had explored a variety of human factors and technical issues--microphone handling, children's reactions to recognizer errors, children's responses

to prompts, the need for adult supervision, proper settings for the recognizer--and we now wanted to examine the effects of microphone quality. We hypothesized that a good quality, noise-cancelling microphone would improve the performance of the recognizer.

RESULTS AND DISCUSSION

Our data on errors address our question about the reliability of the recognizer with young children's speech. In the Phase 1 testing, all but one student had to retrain at least one of the ten vocabulary words. That is, the recognizer could not form a satisfactory template from the initial four utterances, making it necessary for them to give four more. The range of training repetitions needed was from zero (one child) to ten (one child), with a mean of 3.9. The training process proceeded more smoothly in Phase 2 than it had in Phase 1. On the original set of ten vocabulary words the range of training repetitions needed in Phase 2 was from zero (four children) to eight (one child), with a mean of 2.2. This improvement is attributed to modifications in the recognition software, hardware adjustments, and the stabilization of the microphone. In Phase 3 the range of training repetitions was from zero (one child) to nine (one child), with a mean of 3.3. The increased prompting and feedback during Phase 3 training helped students learn how to speak to the computer and to move through the training more quickly, although the amount of retraining required was somewhat higher than in Phase 2.

The use of a high quality, noise-cancelling headset microphone in Phase 4 resulted, with one exception, in an overall improvement in the performance of the recognizer. In training, for example, the new microphone led to a significant reduction in the number of training repeti-

tions.

The one exception was the word "feed," which alone accounted for eight of the twenty-one training repetitions in Phase 4. ("Feed" also presented problems for the recognizer during the story and game portions of the program; see below.) The recognizer's difficulties with this word are probably due to the presence of a long "e" in "feed." A vocalized long "e," such as the one in "feed," produces a low energy sound which is inherently difficult for any microphone to capture accurately. The problem was particularly severe with the microphone we used; it seems that the microphone lacked sufficient sensitivity to handle "feed." A simple, non-technological fix would have been to say "feed" a little more loudly. A more permanent solution would be to adjust the gain setting so that the recognizer becomes more sensitive to low energy sounds.

Excluding "feed," the amount of retraining required by the recognizer in Phase 4 ranged from zero (one child) to four (two children), with a mean of 2.2. (The mean for training repetitions per child when "feed" is included climbs to 3.5, and the range increases from zero to seven.) This information on retraining is contained in Table 1.

[Insert Table 1 about here]

In the story and the game the majority of errors were rejection errors, that is, the recognizer refused to accept a valid utterance of a vocabulary word. In Phase 1 the story rejections accounted for 24 of the 28 total errors. In addition, there were 3 insertions and 1 substitution. In the game, rejections were still the largest category of error, but by a smaller proportion. They accounted for 41 of the 90 total errors, along with 37 insertions and 12 substitutions. In Phases 2 and 3 rejections

continued to be the most common type of error, with the combined rates of substitutions and insertions reduced to 4 percent of Phase 2 errors and 6 percent of Phase 3.

In Phase 4 we collected separate data for error rates in the story, game, and menu sections of the program. The story had only one rejection error: the recognizer refused to accept a valid utterance of the word "feed." Games, as expected, were more problematic: 28 rejection errors were recorded, of which 14 involved the word "feed"; in addition, there were 15 substitution errors, of which 9 were attributable to "feed." The menu--that portion of the program that allows students to move between the various sections of the program by responding with a "yes" or "no"--produced 22 rejection errors which were evenly divided between "yes" and "no"; in addition, there were two substitution errors (see Table 2).

The data on frequency of errors in the story and games show an overall reduction over the course of our testing in the proportion of insertions and substitutions and a corresponding increase in the proportion of rejections. This shift in error frequencies was the expected result of the between-phase adjustments to the recognition software and gain settings, and provides evidence that these adjustments improved the performance of the recognizer. As a general rule, rejection errors do not present a significant problem for the user, so long as the number of errors remains relatively small--say no more than three at one time. The user simply repeats a given word a few times until the recognizer accepts the utterance as valid. Substitution and insertion errors, however, are much more serious. When the recognizer substitutes a different word for the one uttered, or recognizes an extraneous noise as valid, the effect of even a single error is to mislead or confuse the user.

[Insert Table 2 about here]

Though this study was exploratory in nature and not designed to generate precise measurements of recognizer accuracy or comparisons across phases, we did select a small number of participants from Phases 2 and 3 and all the participants in Phase 4 for more detailed analysis of data on recognition accuracy during games--the testing activity that posed the most difficult speech recognition task. Using videotapes, we chose children who represented a range in terms of their total utterances and total errors. Error rates, shown in Table 3, were calculated for each child by dividing the number of each major type of error--rejections and misrecognitions--during a game by that child's total utterances during the game. These error rates should be viewed cautiously within the context of the field conditions in which they were obtained--conditions which differed dramatically from the environment in which the Mark II achieved a 99.3 percent accuracy score (see note 1). Given first-time child users, microphone differences, and a natural noise environment, we might conclude that the Mark II performed surprisingly well for most children. Even those children for whom the recognizer performed less well were able to read more words by the end of the session and, when interviewed, reported that they found the game the "most fun" and that they would like to use the system again at another time.

[Insert Table 3 about here]

We noted several possible explanations for recognizer errors. Some of them may have resulted from environmental background noise. Though the settings in which testing occurred were generally quiet--even compared,

say, to a regular classroom--there were moment to moment variations as announcements were made via loudspeaker or as supervised groups of children passed through the room. These factors were present mainly in the first two days of Phase 1, when the testing area was in the school gymnasium. Subsequent testing took place in a school library and computer room where such interruptions did not occur. Probably more related to speech recognition errors were extraneous verbalizations by participants or their partners. This was particularly evident as a cause in Phase 1 for the much higher rate of insertion errors in the game than in the story. Because many of the children needed assistance with the rules of the game and with strategy, as well as with reading the words, there was a great deal more conversation back and forth between the experimenter and the children and between the children and their partners. The recognizer at times accepted this background conversation as a valid utterance. The addition of prompts and explanations of game rules to the program itself helped to reduce the need for background conversation in Phases 2, 3, and 4. Unlike many speech recognition systems, the Mark II uses an open microphone; that is, it does not require the speaker to use a press-to-talk button or a microphone on/off switch. This makes the system easier to use, but it also requires the system to work harder to distinguish user speech from other speech and noise in the environment.

In some cases, the handling of the microphone appeared to be the cause of problems. On the first day of Phase 1 testing, when a hand-held microphone was used, many children tended to hold the microphone too close or too far away from their mouths or to wave it around, making it difficult for the recognizer to construct usable templates of their speech or to match subsequent utterances once the templates were stored.

Use of a microphone stand on the second day of the first field test reduced these problems and contributed to a decrease in the need for retraining and in certain types of errors during the story and game. In Phase 2, continued use of the microphone stand and an adjustment of the amplification improved performance even more, especially during training.

Some errors may have been due to variability among the utterances of particular children as they tried to accommodate to the system. For example, if children were initially shy and spoke softly to the computer, the recognizer may have formed templates from only the portions of their utterances that were loud enough for it to hear. Later, as students became more confident and talked more loudly, the recognizer may have produced errors because their utterances did not match well against the templates formed earlier. This explanation, for example, may account for the large number of rejection errors involving the word "yes," the first word trained by the students. In Phases 3 and 4 the combination of a headset microphone and gain adjustments kept recognizer performance stable during the story and game while allowing children to speak less loudly than in Phases 1 and 2.

Several human factors emerged as important considerations in speech technology applications with young children. A hand-held microphone may be more than adequate when carefully and consistently used, but our data suggest that four- to six-year-old children, especially when less closely supervised than they were in this experimental setting or when their attention is devoted more to the task at hand than to the way they are holding the microphone, are sufficiently erratic in their handling of the microphone to reduce substantially the recognizer's accuracy.

Use of a microphone stand helped considerably but still required

children to monitor their position relative to it and to learn to speak into it. One of the advantages of the headset microphones used in Phases 3 and 4 was that they required the least conscious attention from participants. A problem with the headset microphones we used, however, was that they were designed for adult heads and adjusted poorly to fit young children. Thus, they sat precariously on some children and were mildly uncomfortable for others. In addition, the particular microphone used in Phase 3 had a tendency to pick up static, which caused the program to crash. After several such episodes on the first day of Phase 3 testing, we solved this problem by using an anti-static spray around the computer. Nevertheless, problems of microphone quality persisted, affecting the performance of the recognizer. The use of a better quality headset microphone in Phase 4 resulted in some improvement, particularly in the story, but recognizer performance continued to be troublesome in the game portion of the program. Some additional improvement can be expected with the use of a headset microphone that is designed for children's heads.

Another human factor of interest was whether the variability in the volume or pronunciation of children's speech would produce problems for the speech recognizer. We found that most children had an initial tendency to speak too softly, but that most quickly became accustomed to the volume level required, especially in Phases 3 and 4 when this level was closer to the one they used in normal conversation. The use of the "too loud" and "too soft" prompts in Phases 3 and 4 helped children to modulate their voices appropriately and to do so with less prompting from the experimenter. Some children, usually those who were having the most difficulty getting the recognizer to accept their utterances, sought to accommodate the machine by altering their pronunciation; they enunciated

more clearly and spoke more loudly or more slowly, all techniques appropriate for a human listener but likely to aggravate the problem for a speech recognition device. Fortunately, this occurred with only a few children. In addition, certain vocabulary words seemed to produce more errors than others. The word "elephant" produced the most errors in Phases 1 and 2, and the second most in Phase 3. "Yes," "feed," and "monkey" also proved difficult for the recognizer. As already noted, "feed" proved especially troublesome in Phase 4.

As we collected our data on machine errors, we also recorded children's reactions to these errors and any possible effects on their performance or their enjoyment of the activity. We found considerable variability on this point. Some children seemed not to mind having to retrain several words and even benefited from the extra opportunities to see the word on the screen. A few became annoyed with having to repeat the same word so many times. This was ameliorated in Phase 2 when the program was modified to take a word that required more than one retraining and move it to the end of the list, rather than train it again for the third consecutive time. The "too loud" and "too soft" prompts in Phases 3 and 4 also helped by letting the children know that the recognizer had not "heard" them and why.

Recognition errors in the story and game were also received differently by different children. Some children were very persistent in their attempts to make the recognizer understand them and would confidently repeat answers that were rejected. Other children waited passively if their first attempt was not accepted. Some children were quite willing to guess, even on unfamiliar words, while others preferred to wait indefinitely for prompts or to rely on the experimenter for assistance.

Another human factor we were concerned about was whether children would be able to understand the speech output needed to deliver prompts, directions, and rewards. In fact, the speech output proved to work very well. Even the lesser quality speech output used in Phases 2, 3, and 4 proved to be well understood in the connected language output portions that instruct students in how to use the machine and how to respond to the tasks presented. Even in the training segment, when single words were initially presented without context, most children had no difficulty understanding the words. A few made minor errors such as hearing "seed" for "feed" or "carrot" for "parrot", but corrected their errors as the feed and parrot graphics appeared on the screen. Only three children persisted in their "mishear" once the graphics were displayed, and had to be told the correct word by the experimenter.

In certain places we found that we had to modify or add speech output prompts to guide students toward patterns of use that would make the recognizer work more reliably or would make the student's experience more satisfying and rewarding. The "too loud" and "too soft" prompts already mentioned are one example. Another is the set of prompts eventually used to introduce and demonstrate the training procedure. Because each word has to be repeated several times during the training, some means is necessary to move children along from one repetition to the next with appropriate pauses between. Initially this was done through "please say" prompts to cue the utterance and changes in the graphic display to reward it. We expected that students would attend strongly to the graphics and begin to anticipate the speech output prompt. When this failed to happen in Phase 1, we inserted delays after the "please say" prompts in Phase 2 to give students an obvious opportunity to override the prompt and use the

graphic reward as a cue that the recognizer was ready for the next repetition. Either because the speech prompts were simply much more salient than the graphics or because the pattern of waiting for the "please say" prompt had already become too firmly established, most children did not spontaneously begin to override the speech prompt. Thus, the delayed prompt, rather than speeding the training process, had the effect of slowing it down. Finally, in Phase 3 we modified the prompt so that it explicitly instructed children to attend to the moving word on the screen and demonstrated how to do so. This explicit approach appears to be necessary, at least with kindergarten and first-grade students.

We also looked at the level of adult support and supervision required to keep students moving along in the program activities. We started in Phase 1 with sparse prompts and directions and found that students tended to rely on the experimenter to tell them when and how to respond. By Phase 3 the number of prompts, directions, and demonstrations had increased in all segments, lessening the need for experimenter supervision. Even by Phases 3 and 4, however, adult supervision was clearly required, not only to make sure children understood what they were expected to do, but also to provide encouragement for them to try again when the recognizer rejected their correct responses or did not respond to them.

This need for adult supervision results from an inherent limitation of the program: unlike the human prompter, the program cannot distinguish between recognizer error and child error. It cannot, depending on the case, encourage a child to repeat the correct word again (explaining, as the experimenter did in our tests, that "sometimes the machine doesn't hear you correctly") or to try a different word.

Certain aspects of children's interaction with program prompts bear on program design and illustrate current constraints of speech technology. Because the speech synthesis chip produces speech that never varies in intonation, recurring lengthy instructions or explanations can prove annoying to users. This effect is compounded by the program's inflexibility: while a human prompter will adjust or interrupt instructions depending on a child's facial or verbal reactions, the program must continue until it has finished what it was programmed to say.

Two ways to deal with this irritation factor are to use speech prompts that are as brief as possible and to include more than one version of a prompt that is to be repeated. For example, the responses to too loud or too soft utterances had two versions which the program used alternately. "Louder, please," alternated with "I can't hear you," and "That's too loud" alternated with "Not so loud, please."

As in their varied responses to program errors, children also differed in their reliance on prompts. Some were reluctant to speak at all and waited for either the program or the experimenter to urge them to respond. Several required the nodded approval of the experimenter before they would respond to the prompts in the program--requiring two levels of encouragement. Other children, by contrast, were impatient with prompts; they frequently spoke while the program was prompting and therefore was not ready to listen to them. The program has only a limited ability to react to such variations in personality style. It will, for instance, repeat prompts if a child does not respond at all, but it cannot adjust the timing of such repetitions for each user, nor can it vary the content of the prompts to suit individual needs.

The program, then, is quite responsive in a "human"-like way to some

of a child's input, but is both inflexible and undiscriminating in other aspects of its interaction. Our observations suggest that most of the children who participated in our study readily adjusted to this combination of program capabilities and limitations.

Finally, we were interested in the instructional potential of speech technology for beginning reading. The data we have thus far collected indicate that such potential does exist. Tables 4, 5, 6, and 7 contain pre- and posttest data on numbers of words learned by students in each phase. In Phase 1 three students learned eight words, four students learned seven words, three students learned six words, two learned five words, two learned four, two learned three, and one student learned two words. Teacher-rated ability level, but not age, appeared to be a predictor of success in learning. Within pairs of students, neither order of participation nor order of posttesting seemed to affect the number of words learned. Students were all in their ninth month of kindergarten; our data suggest that high and average ability readers tended to learn more new words than low readers.

[Insert Table 4 about here]

Again in Phase 2, high and average ability students overall seemed to learn more words among the kindergarteners. Among the high ability kindergarteners one student learned 8 words, one learned 5 words, and one learned 2 words. This group overlapped with the two average students who learned 2 and 4 words respectively. The low ability kindergartener learned no words. Among the first graders the high ability students could already read most of the words. The four low and average ability first-graders learned 8, 6, 6, and 7 words respectively.

[Insert Tables 5 and 6 about here]

In Phase 3 average and high ability children again were most successful, learning in the range of one to eight words. Low ability children in Phase 3 learned no words.

As already noted, Phase 4 students were unrated as to their reading ability. Nonetheless, both in age and classroom experience--all had completed kindergarten and had an average age of about six and a half years--they are comparable to the average and high ability end-of-year kindergarteners of Phase 1 and the low and average ability beginning-of-year first graders in Phase 2. Indeed, an examination of Tables 4, 5, and 7 bears out this comparison: the range of new words learned in Phase 4 was between four and nine, with a mean of 6.2 (excluding one student who already knew all the words). For Phase 1, the range is four to eight new words learned, with a mean of 6.5; for Phase 2 the range is six to eight, with a mean of 6.75.

These results suggest that the likelihood of children's benefiting instructionally from this particular educational application of speech recognition is related to their current reading and reading readiness levels. Specifically, our study indicates that the most effective use of our early reading program may have occurred with children who fit within a relatively well-specified range: end of kindergarten to beginning of first grade. Given that our reading ability levels were supplied by teachers, we conclude that teachers are able to make useful classroom judgments about which children would benefit from participation.

In conclusion, this exploratory study suggests that speech recog-

nition technology holds potential for such educational applications as beginning reading instruction. Perhaps our most significant finding to date is that human factors are an absolutely crucial ingredient in the successful application of speech recognition technology in education. These factors include microphone handling; responses to recognition errors; responses to prompts and rewards; and need for adult supervision. Our data also suggest that gain and reject threshold settings play important roles in recognition accuracy. Finding the proper reject threshold setting involves something of a tradeoff: a setting that decreases rejection errors is likely to create an increase in misrecognition--substitution and insertion--errors. Finally, our research shows that microphone quality is a third important contributor to the system's performance. Further testing is indicated to determine more precisely the kind of microphone that will work most effectively and consistently for young children at the lowest cost. We hypothesize that with optimal human factors conditions, proper gain and reject threshold settings, and an adequate microphone, speech recognition technology can be used effectively in education.

Further testing of a stable group of participants over a period of a few weeks is also needed to determine whether templates would be usable from session to session, as would be expected on the basis of successful applications with adult users. Longer term use would also indicate whether the adult pattern of improved recognizer performance beyond first-time use would occur with children as well. Finally, repeated use would yield evidence as to whether greater familiarity would increase student enjoyment, independence, and learning with the system.

FOOTNOTE

\1/This 99.3 percent accuracy level was obtained in a standard performance test for speech recognition systems. The test is conducted in a sound-treated room, using a high performance, noise-canceling microphone. Training of twenty vocabulary words is accomplished using ten tape-recorded utterances of each word by sixteen speakers, half of whom are first-time users. Speech recognition performance is then evaluated on the basis of sixteen tape-recorded test utterances of the same group of speakers. See Doddington and Schalk (1981) for a detailed description of the standard evaluation procedures.

BIBLIOGRAPHY

Baker, J.M. (1981, Fall). How to achieve recognition: A tutorial/ status report on automatic speech recognition. Speech Technology.

Doddington, G.R., & Schalk, T.B. (1981, September). Speech recognition: Turning theory into practice. Spectrum, pp. 26-32.

Helliwell, J. (1984, June 26). Talking about voice activation. PC Magazine, pp. 171-188.

Hunt, J.A., & Handa, M.K. (1984, June). Speech recognition struggles to life. High Technology, pp. 30-32.

Petre, P. (1985, January 7). Speak, master: Typewriters that take dictation. Fortune, pp. 74-78.

Sandberg-Diment, E. (1984, September 11). Talking back to your computer. New York Times, Sec. C, p. 3.

Sandberg-Diment, E. (1984, September 18). Computer learns its master's voice. New York Times, Sec C, p. 4.

Sandberg-Diment, E. (1984, September 25). Hearing is not so easy either. New York Times, Sec. C, p. 5.

Table 1. Number of retrainings needed to construct satisfactory templates of first ten words.

PHASE 1		
No. Retrainings Needed	No. Children	TOTALS
0	1	0
1	3	3
2	3	6
3	1	3
4	2	8
5	3	15
6	1	6
7	1	7
8	0	0
9	1	9
10	1	10
	17	67
Range: 0-10	Average: 3.9	

PHASE 2		
No. Retrainings Needed	No. Children	TOTALS
0	4	0
1	3	3
2	2	4
3	3	9
4	1	4
5	1	5
6	0	0
7	0	0
8	1	8
	15	33
Range: 0-8	Average: 2.2	

PHASE 3		
No. Retrainings Needed	No. Children	TOTALS
0	1	0
1	1	1
2	4	8
3	0	0
4	1	4
5	1	5
6	1	6
7	0	0
8	0	0
9	1	9
	10	33
Range: 0-9	Average: 3.3	

PHASE 4				
No. Retrainings Needed	No. Children		TOTALS	
	w/ "feed"	w/o "feed"	w/ "feed"	w/o "feed"
0	1	1	0	0
1	1	2	1	2
2	0	0	0	0
3	0	1	0	3
4	2	2	8	8
5	1	0	5	0
6	0	0	0	0
7	1	0	7	0
	6	6	21	13
Range w/ "feed": 0-7 w/o "feed": 0-4			Average w/ "feed": 3.5 w/o "feed": 2.2	

Table 2. Frequency of three kinds of speech recognition errors in games and story.

GAMES						
	No. Games	Rejection Errors No. (%)	Substitution Errors No. (%)	Insertion Errors No. (%)	Total Errors	Mean Errors/ Game
Phase 1	16	41 (46)	12 (13)	37 (41)	90	5.63
Phase 2	20	97 (96)	3 (3)	1 (1)	101	5.05
Phase 3	13	49 (94)	3 (6)	0 (0)	52	4.00
Phase 4	9	28 (65) (w/ "feed")	15 (35)	0 (0)	43	4.80
		14 (61) (w/o "feed")	9 (39)	0 (0)	23	2.50

STORY						
	No. Stories	Rejection Errors No. (%)	Substitution Errors No. (%)	Insertion Errors No. (%)	Total Errors	Mean Errors/ Story
Phase 1	16	24 (86)	1 (3)	3 (11)	28	1.70
Phase 2	12	41 (95)	2 (5)	0 (0)	43	3.58
Phase 3	10	46 (100)	0 (0)	0 (0)	46	4.60
Phase 4	6	1 (100) (w/ "feed")	0 (0)	0 (0)	1	0.17

MENU						
	No. Utterances	Rejection Errors No. (%)	Substitution Errors No. (%)	Insertion Errors No. (%)	Total Errors	Mean Errors/ Utterance
Phase 4	93	22 (92)	2 (8)	0 (0)	24	0.26

Table 3. Error rates during games for selected participants in Phase 2, 3, and 4.

Child	No. Utterances	No. Rejection Errors	Percent Rejection Errors	No. Misrecognition Errors	Percent Misrecognition Errors
PHASE 2					
1	34	9	26	0	0
2	20	4	20	0	0
3	61	9	15	1	2
PHASE 3					
4	41	5	12	0	0
5	23	7	30	1	4
6	8	2	25	0	0
PHASE 4					
7	37	14*	38	1	3
8	36	8**	22	8***	22
9	42	2	5	5****	12
10	43	1	2	1	2
11	13	3	23	0	0
12	10	0	0	0	0

* 12 of 14 errors involved the word "feed"
** 2 of 8 errors involved the word "feed"
*** 6 of 8 errors involved the word "feed"
**** 2 of 5 errors involved the word "feed"

Table 4. Number of words learned in Phase 1 (out of ten).

PHASE 1 Kindergarteners--9th month				
Teacher Rating of Ability	Mean Age Yrs-Mos	Mean Pretest Score	Mean Posttest Score	Mean Words Learned (Range)
Low (n=5)	5-7	0.0	3.6	3.6 (2-6)
Average (n=5)	6-0	0.0	6.4	6.4 (5-8)
High (n=7)	5-8	1.1	7.7	6.6 (4-8)

Table 5. Number of words learned in Phase 2 (out of ten).

PHASE 2 Kindergarteners--3rd month				
Teacher Rating of Ability	Mean Age Yrs-Mos	Mean Pretest Score	Mean Posttest Score	Mean Words Learned (Range)
Low (n=1)	5-2	0.0	0.0	0.0
Average (n=2)	4-2	0.0	3.0	3.0 (2-4)
High (n=3)	5-3	1.3	5.0	3.7 (2-8)
Unrated (n=1)	5-4	0.0	2.0	2.0
First Graders--3rd month				
Low (n=2)	6-6	0.5	7.5	7.0 (6-8)
Average (n=2)	6-8	2.0	8.5	6.5 (6-7)
High (n=4)	6-6	8.5	10.0	1.5 (0-4)

Table 6. Number of words learned in Phase 3 (out of ten).

PHASE 3 Kindergarteners--4th month				
Teacher Rating of Ability	Mean Age Yrs-Mos	Mean Pretest Score	Mean Posttest Score	Mean Words Learned (Range)
Low (n=3)	5-4	2.7	2.3	0.0
Average (n=2)	5-0	0.5	2.5	2.0 (1-3)
High (n=5)	5-3	3.4	7.4	4.0 (2-8)

Table 7. Number of words learned in Phase 4 (out of ten).

PHASE 4 Summer before First Grade				
Teacher Rating of Ability	Mean Age Yrs-Mos	Mean Pretest Score	Mean Posttest Score	Mean Words Learned (Range)
Unrated (n=5*)	6-4	1.2	7.4	6.2 (4-9)

* Table excludes data on one student who already knew all ten words at pretest.